What day of the week are there the most searches for [hangover]?

1. Sunday
2. Monday
3. Tuesday
4. Wednesday
5. Thursday
6. Friday
7. Saturday
Searches for [hangover]

Interest over time

The number 100 represents the peak search volume

Regional interest

Worldwide > United States

Related terms

cure hangover 100
hangover cures 65
hangover remedies 35

Map of U.S. showing regional interest in hangover-related searches.
Searches for [hangover] and [vodka]
Looking for gifts when single

1. [gift for boyfriend]
2. [gift for girlfriend]
Looking for gifts when married

1. [gift for husband]
2. [gift for wife]
Want to use Google Trends data to nowcast economic series
- unemployment may be predicted by “job search” queries
- auto purchases may be predicted by “vehicle shopping” queries
- often a contemporaneous relationship, hence “nowcasting”
- useful due to reporting lags and revisions

Fat regression problem: there are many more predictors than observations

Millions of queries, hundreds of categories
- number of observations $\sim 100$ for monthly economic data
- number of predictors $\sim 150$ for “economic” categories in Trends

How do we choose which variables to include?
Example: unemployment

- Sometimes Google Correlate works
- Load in: data on initial claims for unemployment benefits
- Returns: 100 queries, including [sign up for unemployment]
Build a simple AR model

- Use deseasonalized initial claims ($y_t$)
- Use deseasonalized, detrended searches for [sign up for unemployment] ($x_t$)

**Base:**  
$$y_t = a_0 + a_1 y_{t-1} + e_t$$

**Regr:**  
$$y_t = a_0 + a_1 y_{t-1} + bx_t + e_t$$

- Estimate regressions using rolling window
- One-step-ahead MAE during recession is about 8.7% lower when [sign up for unemployment] query is included
But sometimes simple correlation doesn’t work

User uploaded activity for **US Auto Sales NSA** and United States Web Search activity for **Indian restaurants**

(r=0.7195)

Hint: Drag to Zoom, and then correlate over that time only.
How to avoid spurious correlation?

- Control for trend and seasonality
  - Build a model for the *predictable* (trend + seasonality) part of time series
  - In time series this is called *whitening* or *prewhitening*
  - Find regressors that predict the *residuals* after removing trend and seasonality

- How to choose regressors?
  - Simple correlation is too limited
  - Human judgment doesn’t scale
Some approaches to variable selection

- Human judgment: what we mostly do
- Significance testing: forward and backward stepwise regression
- Complexity criteria: AIC, BIC, etc
- Dimensionality reduction: principle component, factor models, partial least squares
- Machine learning: Penalized regression, lasso, LARS, ridge regression, elastic net
Our approach

- Bayesian Structural Time Series (BSTS)
  - Decompose time series into trend + seasonality + regression
  - Use Kalman filter for trend + seasonality (whiten time series)
  - Spike and slab regression for variable selection
  - Estimate via Markov Chain Monte Carlo simulation of posterior distribution
  - Bayesian model averaging for final forecast
How BSTS helps reduce overfitting

- Kalman filter used to whiten the series
  - Remove common seasonality and trend, regressors chosen to predict residuals
  - Estimation of (seasonality, trend, regression) is simultaneous
  - Same spirit as Granger causality
- Overfitting due to spurious correlation with regressors
  - Remove “one time” events (can be automated)
  - Apply human judgment
- Overfitting due to many regressors
  - Informative prior to suggest likely number of regressions or regressor categories
  - Bayesian model averaging over many small regressions (“ensemble estimation”)
Consider classic time series model with *constant* level, linear time trend, and regressors

\[ y_t = \mu + bt + \beta x_t + e_t \]

“Local linear trend” is a stochastic generalization of this

- Observation: \( y_t = \mu_t + z_t + e_{1t} = \text{level} + \text{regression} \)
- State 1: \( \mu_t = \mu_{t-1} + b_{t-1} + e_{2t} = \text{random walk} + \text{trend} \)
- State 2: \( b_t = b_{t-1} + e_{3t} = \text{random walk for trend} \)
- State 3: \( z_t = \beta x_t = \text{regression} \)

Parameters to estimate: regression coefficients \( \beta \) and variances of \( (e_{it}) \) for \( i = 1, \ldots, 3 \)

Use these variances to construct optimal Kalman forecast:

\[ \hat{y}_t = \hat{y}_{t-1} + k_t \times (y_{t-1} - \hat{y}_{t-1}) + x_t \beta \]

\( k_t \) depends on the estimated variances
Intuition for Kalman filter

- Consider simple case without regressors and trend
  - Observation equation: $y_t = \mu_t + e_{1t}$
  - State equation: $\mu_t = \mu_{t-1} + e_{2t}$

- Two extreme cases
  - $e_{2t} = 0$ is constant mean model where best estimate is sample average through $t - 1$: $\bar{y}_{t-1} = \sum_{s=1}^{t-1} y_s$
  - $e_{1t} = 0$ is random walk where best estimate is current value $y_{t-1}$

- For general case take weighted average of current and past observations, where weight depends on estimated variances
Nice features of Kalman approach

- No problem with unit roots or other kinds of nonstationarity
- No problem with missing observations
- No problem with mixed frequency
- No differencing or identification stage (easy to automate)
- Nice Bayesian interpretation
- Easy to compute estimates (particularly in Bayesian case)
- Nice interpretation of structural components
- Easy to add seasonality
- Good forecast performance
Spike and slab regression for variable choice

- **Spike**
  - Define vector $\gamma$ that indicates variable inclusion
  - $\gamma_i = 1$ if variable $i$ has non-zero coefficient in regression, 0 otherwise
  - Bernoulli prior distribution, $p(\gamma)$, for $\gamma$
  - Can use an informative prior; e.g., expected number of predictors

- **Slab**
  - Conditional on being in regression ($\gamma_i = 1$) put a (weak) prior on $\beta_i$, $p(\beta|\gamma)$.
  - Estimate posterior distribution of $(\gamma, \beta)$ using MCMC
Bayesian model averaging

- We simulate draws from posterior using MCMC
- Each draw has a set of variables in the regression ($\gamma$) and a set of regression coefficients ($\beta$)
- Make a forecast of $y_t$ using these coefficients
- This gives the posterior forecast distribution for $y_t$
- Can take average over all the forecasts for final prediction
- Can take average over draws of $\gamma$ to see which predictors have high probability of being in regression
Torture test simulation for BSTS

- Pick $k = 3$ categories (out of 150) and their associated time series
- Construct artificial time series = sum of these $k +$ noise
- See if BSTS picks the right categories
  - 0 noise = perfect
  - 5% noise = perfect
  - 10% noise = misses one, but still does good forecast
  - performance deteriorates for higher noise levels
  - ... but it degrades gracefully
Example of torture test

- **SD = .05**
  - Arts_and_Entertainment
  - Fitness
  - Card_Games

- **SD = .10**
  - Programming
  - Arts_and_Entertainment
  - Fitness
  - Card_Games

- **SD = .20**
  - Arts_and_Entertainment
  - Fitness
  - Card_Games

- **SD = .30**
  - Arts_and_Entertainment
  - Programming
  - Fitness
  - Card_Games

- **SD = .40**
  - Fitness
  - Card_Games

- **SD = .50**
  - Arts_and_Entertainment
  - Programming
  - Fitness
  - Card_Games
Example 1: Consumer Sentiment

- Monthly UM Consumer sentiment from Jan 2004 to Apr 2012 ($n = 100$)
- Google Insights for Search categories related to economics ($k = 150$)
- No compelling intuition about what predictors should be
Consumer sentiment as leading indicator

- Leading indicator of retail sales in last recession
Google Insights for Search categories related to economics ($k = 150$)
- Deseasonalize predictors using R command `stl`
- Detrend predictors using simple linear regression
- Let `bsts` choose predictors
UM Consumer Sentiment Predictors

Probability of inclusion

- Financial.Planning
- Investing
- Business.News
- Search_ENGINES
- Energy.UTILITIES
- Hybrid.Alternative.Vehicles

Inclusion Probability
Posterior distribution of one-step ahead forecast
State decomposition

Recall observation equation:

\[ y_t = \mu_t + x_t \beta + e_{1t} \]

We can plot the posterior distribution of each of these components. The regression component can be further expanded

\[ y_t = \mu_t + x_{1t} \beta_1 + \cdots + x_{pt} \beta_p + e_{1t} \]

Natural to order predictors by probability of inclusion and look at cumulative plot.
Trend and regression decomposition

```
trend

regression

distribution

time

2004 2008 2012

-100 -50 0 50 100 150

distribution

time

2004 2008 2012

-100 -50 0 50 100 150
```
Trend

1. trend (mae=5.6656)
2. add Financial.Planning (mae=4.8529)
4. add Investing (mae=3.3511)
5. add Search.Engines (mae=3.2748)
Example 2: gun sales

Use FBI’s National Instant Criminal Background Check
Google Correlate Results

- [stack on] has highest correlation
- [gun shops] is chosen by bsts
- Regression model gives 11% improvement in one-step ahead MAE
Trend

1. trend (mae=0.49947)
2. add seasonal (mae=0.33654)
Gun Shops

3. add gun.shops (mae=0.15333)
Google Trends predictors

- 586 Google Trends verticals, deseasonalized and detrended
- 107 monthly observations

<table>
<thead>
<tr>
<th>Category</th>
<th>mean</th>
<th>inc.prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recreation::Outdoors::Hunting:and:Shooting</td>
<td>1,056,208</td>
<td>0.97</td>
</tr>
<tr>
<td>Travel::Adventure:Travel</td>
<td>-84,467</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table: Google Trends predictors for NICS checks.
1. trend (mae=130270)
2. add seasonal (mae=61094)
3. add recreation_shooting (mae=43128)
Searches for [gun shop]

Interest over time

Regional interest
Worldwide > United States

Related terms
buds gun shop 100
the gun shop 20
gun shops 15
bass pro 15
bass pro shop 15
pawn shop 15
gun store 10
buds guns 5
online gun shop 5
bills gun shop 5
Can use prior to improve estimate of trend component
  - Google data starts in 2004, only one recession
  - Can estimate parameters of trend model with no regressors
  - Use this as prior for estimate of trend in estimation period

Can use prior to influence variable choice in regression
  - Influence the expected number of variables in regression (parsimony)
  - Give higher weight to certain verticals (e.g., economics related)
  - Exclude obvious spurious correlation (e.g., pop song titles)
New Homes Sold in the US

New Homes Sold in the United States (HSN1FNSA)
Source: U.S. Department of Commerce: Census Bureau

Shaded areas indicate US recessions.
2013 research.stlouisfed.org
Run correlate

Google correlate

Correlated with HSN1FNSA

0.819 tahitian noni juice
0.809 exhaust sound
0.902 trderonline.com
0.759 www.lbb.com

0.9788 80/20 mortgage
0.9782 appreciation rate
0.9778 home appreciation
0.9759 help-u-sell
0.9759 planned community
0.9758 new home builder

User uploaded activity for HSN1FNSA and United States Web Search activity for 80/20 mortgage (r=0.9786)

Hint: Drag to Zoom, and then correlate over that time only.
BSTS variable selection

With all correlates

- appreciation.rate
- oldies.lyrics
- real.estate.purchase
- irs.1031
- www.mail2web.com
- century.21.realtors
- selling.real.estate

Inclusion Probability
Eliminate spurious correlates
State decomposition
Predictors

![Graph with labeled predictors:]

- appreciation.rate: 0.89
- irs.1031: 0.45
- century.21.realtors: 0.35
- real.estate.appraisal: 0.2
- estate.appraisal: 0.17
- real.estate.purchase: 0.15
- X80.20.mortgage: 0.13
- selling.real.estate: 0.12
Seasonal

2. add seasonal (mae=0.47767)
3. add appreciation.rate (mae=0.2241)
5. add century.21.realtors (mae=0.077138)
Real estate appraisal
Estate appraisal

7. add estate.appraisal (mae=0.16587)
Real estate purchase

8. add real_estate_purchase (mae=0.13757)
9. add X80.20.mortgage (mae=0.11207)
Future work

- Mixed frequency forecasting — done
- Fat tail distributions – underway
- Parallel MCMC – underway
- Panel data
- Automate the whole thing – goal